LATENT RULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Rules are reusable abstract relations or constraints across different contexts. Inferring and applying rules to unseen cases across various tasks remains challenging for specialized models and Large Vision-Language Models (LVLMs). We propose **Latent Rule Reasoning (LRR)**, which integrates sequential rule induction and execution within an LVLM's continuous latent space. LRR induces rule representations from visual context and applies them under query conditions to form target-state representations for answer generation. We further introduce **Latent Rule Self-Distillation (LRSD)**, in which the same LVLM uses complete visual observations available only during training to guide rule induction from inference-time inputs. Training first grounds rule representations and develops target prediction using rule-annotated examples, then incorporates cross-view rule consistency to jointly learn from examples with and without explicit rule annotations while retaining answer supervision. Our experiments reveal that LRR achieves an improvement of 10.70% on VisuLogic and a gain of 6.75% on SpatialEval, demonstrating the effectiveness of latent rule reasoning on various in-distribution, out-of-distribution, and general visual understanding tasks.

1 INTRODUCTION

Rules are abstract relations or constraints that can be reused across different contexts (Marcus et al., 1999; Tenenbaum et al., 2011; Kemp & Tenenbaum, 2008). In abstract reasoning, rules capture attribute-level change patterns and compositional relationships across panels (Carpenter et al., 1990; Raven & Court, 1938). In physical dynamics prediction, rules govern how objects move and interact, determining trajectories under gravity and changes in velocity after collisions (Battaglia et al., 2016; McCloskey, 1983). In 3D scene, rules manifest as geometric constraints linking spatial structure to visual observations, requiring consistency across different views of a static 3D scene (Hartley et al., 2003). Across these scenarios, understanding rules enables AI systems to extract reusable structure from limited observations and apply it to infer unobserved states and outcomes. Learning to identify and apply rules is therefore an important foundation for developing reasoning capabilities that generalize beyond observed examples.

Specialized models learn rules through dedicated mechanisms, some of which are tied to particular task structures, for example, organizing rule induction around ordered rows or columns of a matrix rely on the layout assumptions (Tenenbaum et al., 2011; Kemp & Tenenbaum, 2008). These designs provide useful inductive biases, but can limit transfer to other tasks (Zhang et al., 2021a; Shi et al., 2021). Advances in Large Language Models (LLMs) (Bai et al., 2023; Dubey et al., 2024) have driven the development of Large Vision-Language Models (LVLMs) (Bai et al., 2025b; Wang et al., 2025a; Liu et al., 2024a), which combine visual perception with rich linguistic and world knowledge. However, previous evaluations show that the tested models perform relatively well on basic perception yet struggle to identify rules and handle logical dependencies across reasoning stages (Ahrabian et al., 2024; Anis et al., 2025). Moving toward more general rule discovery, CRD leverages LVLM priors to extract visual attributes and models their variation through learnable external rule functions, but rule inference remains dependent on extraction quality, with no direct feedback from rule learning to refine the LVLM's internal representations (Shi et al., 2026b). Integrating the rule understanding ability into LVLMs, enabling them to leverage perception to learn reusable visual relations and apply them to predict target states, remains an important challenge.

Human cognition offers a useful perspective on this challenge, as the ability to discover rules and apply them in unfamiliar situations is central to fluid intelligence. Classic studies of Raven's Pro-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

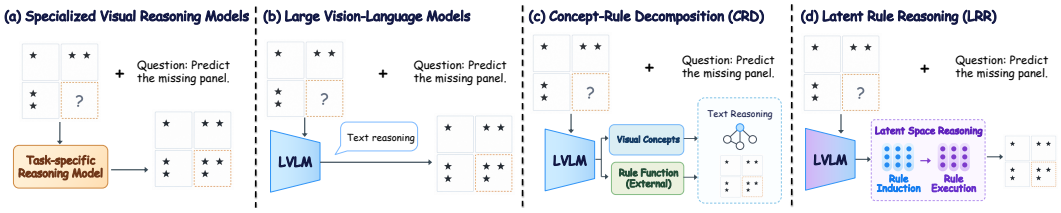

Figure 1: **Illustrative comparison of latent rule reasoning.** LRR integrates rule induction and execution within the LVLM's continuous latent space to predict target states.

gressive Matrices describe how people encode visual attributes, induce relational rules, and apply them through analogy to infer missing information (Carpenter et al., 1990). Such reasoning need not depend on explicit verbalization, consistent with evidence that complex thought can proceed without language (Fedorenko et al., 2024). Neuroscientific research further suggests that cognitive processes underlying inference rely on abstract internal representations of latent task structure (Courellis et al., 2024; Liu et al., 2019). These findings motivate modeling rule understanding as reasoning over latent representations. From a functional perspective, we characterize this process through two complementary capacities. *Rule induction* extracts abstract relations from concrete observations, while *rule execution* applies these relations to query conditions to predict the corresponding target state. They connect observed evidence to underlying rules and their consequences, motivating learning objectives that constrain both rule representations and the prediction of the outcomes they imply.

Building on this perspective, we propose **Latent Rule Reasoning (LRR)**, which integrates rule induction and execution within an LVLM's continuous latent space. LRR first encodes attribute relations from the visual context into a rule representation, then applies the inferred rules under query conditions to form an execution representation of the predicted target state. Together, these intermediate representations support final answer generation. To train LRR, we introduce *Latent Rule Self-Distillation* (LRSD), which leverages complete visual observations available only during training to guide rule induction. We construct a complete view by restoring the ground truth target at the query position. The same LVLM provides rule guidance from this view, while the student learns from the input available at inference. Training initially uses examples with rule annotations to ground rule representations and develop target prediction capabilities. It then incorporates rule consistency supervision between the complete and inference views, enabling joint learning from examples with and without explicit rule annotations. Our contributions are summarized as follows:

- We propose LRR, a framework that structures reasoning in the continuous latent space of LVLMs into sequential rule induction and rule execution processes. LRR first infers rule representations that capture attribute relations in the context, then execute them with query conditions to form representations of the predicted target state to support answer generation.
- We introduce LRSD, a self-distillation mechanism that transfers rule guidance from complete visual observations available during training to the student's reasoning under inference conditions. Through rule consistency learning, it supports joint training on examples with and without explicit rule annotations while retaining answer supervision.
- Our experiments demonstrate that LRR outperforms both specialized models and LVLM-based methods on in-distribution, out-of-distribution and general benchmarks.

2 Related Work

Rule Learning Methods. Early work investigates spatial regularities in structured visual inputs through hierarchical Bayesian inference (Tenenbaum et al., 2011; Kemp & Tenenbaum, 2008). Related approaches to character modeling encode compositional rules over strokes and apply them to generate new characters (Lake et al., 2011; 2015; Shi et al., 2025). In abstract visual reasoning, methods based on hand-crafted or learned feature representations infer relations among visual elements to solve reasoning problems (Lovett & Forbus, 2017; Little et al., 2012; Lovett et al., 2010). Subsequent approaches model visual rules more explicitly by combining latent encodings with probabilistic induction or adversarial learning (Pekar et al., 2020; Zhang et al., 2021a;b; Shi et al., 2024; 2026a). Another line of work represents rules as latent functions governing changes in visual at-

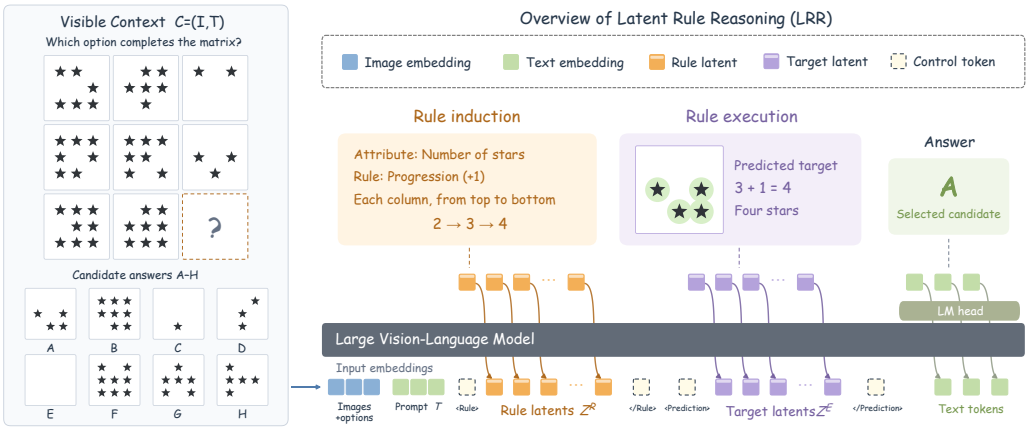

Figure 2: **Overview of Latent Rule Reasoning (LRR).** LRR first encodes the given visible context into rule latents, then excute the rules under query conditions to form target latents for answer generation. In both stages, hidden states are feed back directly as input embeddings, keeping intermediate reasoning within the LVLM's continuous latent space.

tributes, connecting rule inference with the prediction of unobserved states (Shi et al., 2021; 2023). However, many formulations incorporate predefined rule forms or structural assumptions tailored to particular tasks, which can constrain their applicability across problem settings. LRR integrate rule induction and execution into the continuous latent reasoning process of a general LVLM that ground latent representations in rule semantics and their predicted visual consequences.

Large Vision-Language Models. Large Vision-Language Models (LVLMs) extend language modeling to visual inputs by connecting pretrained visual encoders with large language models, enabling visual question answering and instruction following (Liu et al., 2023; Li et al., 2023; Chen et al., 2025). Subsequent work improves visual understanding through higher input resolutions, stronger vision-language alignment, and more effective use of visual features (Liu et al., 2024a; Chen et al., 2024c; Yin et al., 2025; Yu et al., 2025; Liang et al., 2025; 2026). Models such as InternVL3.5 and Qwen2.5-VL further advance multimodal understanding and reasoning through architectural and training improvements (Wang et al., 2025a; Bai et al., 2025b). Building on these capabilities, SpatialVLM and VisuoThink explore dedicated spatial supervision and multimodal reasoning procedures to enhance visual reasoning (Chen et al., 2024a; Wang et al., 2025c). Nevertheless, evaluations have revealed shortcomings in visual grounding and abstract rule induction, indicating that broad multimodal capabilities do not necessarily translate into reliable rule understanding (Tong et al., 2024; Ahrabian et al., 2024). To connect LVLM priors with explicit rule modeling, CRD (Shi et al., 2026b) uses an LVLM to extract visual attributes and models their spatial variation through Gaussian process rule functions, which guide an iterative sampling procedure. While this provides an account of visual regularities, its rule functions are modeled through a probabilistic module external to the LVLM. This separation makes rule inference dependent on the accuracy of explicitly extracted attributes, without allowing rule learning to directly refine internal representations. Inspired by human cognition, LRR internalizes rule induction and execution into the LVLM's continuous latent space.

3 METHOD

We propose a Latent Rule Reasoning (LRR) framework for LVLMs, together with Latent Rule Self-Distillation (LRSD), which is an adaptation of on-policy self-distillation for training. LRR organizes visual reasoning into two stages. The rule induction stage infers rule representations from the visible context, and the rule execution stage applies these rules to the query conditions to predict a target state. The resulting representations are then used to generate the answer. We assign distinct supervision to these stages: reconstructing the complete rule set provides semantic supervision for rule induction, while matching the visual representation of the true target supervises rule execution. LRSD uses a rule teacher to guide the student at the context prefixes that the student actually visits.

3.1 Problem Formulation

A training example consists of a panel image I, a textual prompt T, an answer A, and an optional rule set $R = \{\rho_1, \ldots, \rho_J\}$. We formulate visual reasoning as inferring a target panel U at the query position from the observed context $C = (I, T)$. This formulation is motivated by a simple principle: understanding the rules underlying the observed panels should enable the model to predict the properties of the missing target. For multiple-choice problems, the model must infer which candidate satisfies the underlying rules. Each rule ρ_j specifies a relation bound to a particular attribute, for example, the color of an object follows a `progressive` change pattern.

3.2 Latent Rule Reasoning

LRR introduces two continuous latent sequences into the language backbone of a LVLM:

$$Z^R = (z_1^R, \ldots, z_M^R) \in \mathbb{R}^{M\times d}, \quad Z^E = (z_1^E, \ldots, z_K^E) \in \mathbb{R}^{K\times d}, \tag{1}$$

where d is the input embedding dimension of the language model. The complete input-output sequence is organized as

$$Y = \big[I, T, \texttt{<Rule>}, Z^R, \texttt{</Rule>}, \texttt{<Prediction>}, Z^E, \texttt{</Prediction>}, A\big]. \tag{2}$$

Within each continuous segment, the model directly feeds its final-layer hidden state back as the next input embedding, and segment boundaries use the ordinary embeddings of the control tokens. The LVLM retains its original positional encoding, with positions advancing at every step. We use $M = K = 8$, giving a total budget of 16 continuous latent steps. We describe the reasoning process with the conditional generative model

$$p(A, z \mid C) = p(Z^R \mid C)\, p(Z^E \mid Z^R, C)\, p(A \mid C, Z^R, Z^E) \tag{3}$$

where $z = (Z^R, Z^E)$. This factorization specifies rule induction followed by rule execution, while allowing both stages to access the visible context. The latent priors are autoregressive Gaussians:

$$\begin{aligned} p(Z^R \mid C) &= \prod_{m=1}^{M} \mathcal{N}\left(Z_m^R; \mathrm{LVLM}(C, Z_{<m}^R), \tau_R^2 I\right), \\ p(Z^E \mid Z^R, C) &= \prod_{k=1}^{K} \mathcal{N}\left(Z_k^E; \mathrm{LVLM}(C, Z^R, Z_{<k}^E), \tau_E^2 I\right). \end{aligned} \tag{4}$$

The means are the corresponding final-layer hidden states. The fixed variances τ_R^2 and τ_E^2 specify the scale of uncertainty, which are similar to the temperature in LLMs. The answer distribution $p(A \mid C, Z^R, Z^E)$ is autoregressive for textual answers. Our deterministic implementation feeds back the Gaussian means directly, and it does not set these variances to zero.

3.3 Variational Inference and Parameter Learning of Latent Rules

The priors in Equation 4 define the *student*: they describe how the LVLM induces rules and predicts a target using only information available at inference. During training, we introduce a variational distribution that uses privileged visual information to provide *teacher* guidance:

$$q(z \mid C, U) = q(Z^R \mid C^+)\, q(Z^E \mid U). \tag{5}$$

The rule teacher is the same LVLM conditioned on the complete view. The execution teacher is defined by the frozen visual representation of the true target. Their purpose is to transfer information about the underlying rules and target state to the student. The factorization in Equation 5 is a choice of variational family, where the two latent sequences are regarded independent in the generative model under the complete view. Specifically, we construct the rule posterior as

$$\begin{aligned} q(Z^R \mid C^+) &= \prod_{m=1}^{M} \mathcal{N}\left(Z_m^R; \mathrm{LVLM}(C^+, Z_{<m}^R), \tau_R^2 I\right), \\ H^U = \mathrm{SG}\left(\mathrm{Pruner}(\mathrm{VE}(U))\right) &\in \mathbb{R}^{K\times d}, \quad q(Z^E \mid U) = \prod_{k=1}^{K} \mathcal{N}\left(Z_k^E; H_k^U, \tau_E^2 I\right). \end{aligned} \tag{6}$$

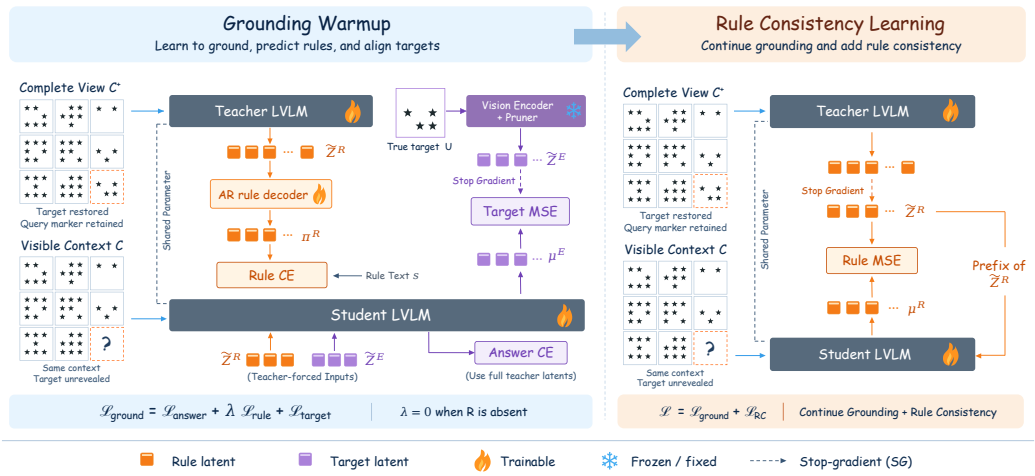

Figure 3: **Overview of Latent Rule Self-Distillation (LRSD).** The teacher and student share LVLM parameters but receive the complete view and visible context, respectively. Grounding warmup combines answer supervision, rule reconstruction, and target alignment with fixed visual features from the true target. Rule consistency learning additionally matches student rule predictions to detached teacher latents using teacher latent prefixes.

Here, VE is a frozen visual encoder, Purner is a fixed token pruner, and SG denotes stop-gradient operation. The target pipeline remains fixed throughout training.

To make the rules recoverable from the learned latent Z^R, we introduce an autoregressive text decoder $p(R \mid Z^R)$. This notation denotes the probability of the serialized complete rule set. Let $\lambda \geq 0$ be the effective rule weight for an example, with $\lambda = 0$ when its rule annotation is unavailable. We define the weighted evidence

$$\mathcal{Z}_\lambda(C, A, R) = \int p(R \mid Z^R)^\lambda \, p(A \mid C, z) \, p(z \mid C) \, \mathrm{d}z. \tag{7}$$

For unannotated examples, the rule factor is omitted. With observed rules and $\lambda = 1$, this quantity is the joint conditional evidence of the answer and rule text, and other positive weights give a weighted evidence. Writing $q = q(z \mid C, U)$ for readability, Jensen's inequality gives

$$\begin{aligned} \log \mathcal{Z}_\lambda &= \log \mathbb{E}_q \left[\frac{p(R \mid Z^R)^\lambda p(A \mid C, z) p_\theta(z \mid C)}{q(z \mid C, U)} \right] \\ &\geq \mathbb{E}_q \left[\log p(A \mid C, z) \right] + \lambda \mathbb{E}_q \left[\log p(R \mid Z^R) \right] - D_{\mathrm{KL}} \left(q(z \mid C, U) \,\|\, p(z \mid C) \right). \end{aligned} \tag{8}$$

Expanding the joint KL using the prior and posterior factorizations yields four terms:

$$\begin{aligned} \mathcal{L}_{\mathrm{ELBO}} = & \underbrace{\mathbb{E}_q \left[\log p(A \mid C, Z^R, Z^E) \right]}_{\text{Answer Prediction Term } \mathcal{J}_{\mathrm{AP}}} + \lambda \underbrace{\mathbb{E}_{q(Z^R \mid C^+)} \left[\log p(R \mid Z^R) \right]}_{\text{Rule Prediction Term } \mathcal{J}_{\mathrm{RP}}} \\ & - \underbrace{D_{\mathrm{KL}} \left(q(Z^R \mid C^+) \,\|\, p(Z^R \mid C) \right)}_{\text{Rule Consistency Term } \mathcal{K}_{\mathrm{RC}}} - \underbrace{\mathbb{E}_{q(Z^R \mid C^+)} \left[D_{\mathrm{KL}} \left(q(Z^E \mid U) \,\|\, p(Z^E \mid Z^R, C) \right) \right]}_{\text{Target Prediction Term } \mathcal{K}_{\mathrm{TP}}}. \end{aligned} \tag{9}$$

The aforementioned subterms are described as follow.

- **Answer Prediction Term.** $\mathcal{J}_{\mathrm{AP}}$ increases the probability of the correct answer under the posterior latents, coupling rule induction and execution to the final reasoning task. The negative $\mathcal{J}_{\mathrm{AP}}$ is the expected answer negative log-likelihood, or the expected candidate cross-entropy for a multiple-choice classifier.
- **Rule Prediction Term.** $\mathcal{J}_{\mathrm{RP}}$ reconstructs all annotated rules from Z^R. The decoder receives only Z^R as its example-specific conditioning representation; ground-truth rule prefixes are used for standard autoregressive teacher forcing. Reconstruction encourages the rule latents to retain attribute-rule bindings.

Algorithm 1 One training iteration of LRSD

Input: Trainable parameters θ; one training sample (C, A, R); Latent budgets $M = K = 8$.

1. Zero gradients for the trainable parameters.
2. Construct C^+ by using the answer A and retaining the query marker.
3. $\widetilde{Z}^E \leftarrow \mathrm{SG}\left(\mathrm{Pruner}(\mathrm{VE}(U))\right)$.
4. $\widetilde{Z}^R \leftarrow \mathrm{TeacherRollout}(C^+; M, K)$.
5. Compute $\mathcal{L} \leftarrow \mathcal{L}_{\text{ground}}$ through Equation 11.
6. **if** the phase is Rule Consistency Learning **then**
   $\widetilde{Z}^R \leftarrow \mathrm{SG}(\widetilde{Z}^R)$
   Compute $\mathcal{L} \leftarrow \mathcal{L} + \sum_{m=1}^{M} \|\mathrm{LVLM}(C, \widetilde{Z}^R_{<m}) - \widetilde{Z}^R_m\|_2^2$ using Equation equation 12.
7. **end if**
8. Backpropagate $\mathcal{L}$ and update θ.

- **Rule Consistency Term.** $\mathcal{K}_{\mathrm{RC}}$ matches the context-only rule prior to the complete-view rule posterior. The posterior can use the restored target to infer rules from a more complete observation, whereas the prior must infer them from the visible context. Their KL therefore provides a distributional teacher-student constraint for rule induction.
- **Target Prediction Term.** $\mathcal{K}_{\mathrm{TP}}$ matches the target predicted through rule execution to the true target features, which constrains the result of rule execution to represent the query target.

3.4 Latent Rule Self-Distillation

LRSD adapts the student-prefix matching principle of on-policy self-distillation to continuous rule representations. The variational formulation motivates its four supervision components, and the practical algorithm uses deterministic teacher rule targets to supervise student rollouts. The model first takes C^+ as input and generates latent tokens that are used as teacher targets:

$$\widetilde{Z}^R \sim q(Z^R \mid C^+), \quad \widetilde{Z}^E \sim q(Z^E \mid U). \tag{10}$$

We adopt deterministic sampling throughout this process by setting the temperature to zero.

Rule Grounding. We restrict this stage to samples that contain both complete rule annotations and valid target panels, while disabling the rule-consistency objective. The grounding loss is

$$\mathcal{L}_{\text{ground}} = \sum_{l=1}^{L_A} \mathrm{CE}\left(\pi_l^A, A\right) + \lambda \sum_{l=1}^{L_R} \mathrm{CE}\left(\pi_l^R, s_l\right) + \left\| \mathrm{LVLM}(C, \widetilde{Z}^R, \widetilde{Z}^E_{<k}) - \widetilde{Z}^E_k \right\|_2^2. \tag{11}$$

Since the latent trajectories are sampled from the teacher posterior, relying solely on posterior supervision would leave the prior without a direct learning signal. We therefore optimize the model through a fully observed posterior, in which complete observations are used to construct posterior trajectories for supervision. This warmup stage is similar to Supervised Fine-Tuning (SFT).

Rule Consistency Learning. We introduce the rule consistency term to the LRSD objective:

$$\mathcal{L}_{\text{LRSD}} = \mathcal{L}_{\text{ground}} + \sum_{m=1}^{M} \left\| \mathrm{LVLM}(C, \widetilde{Z}^R_{<m}) - \widetilde{Z}^R_m \right\|_2^2. \tag{12}$$

The intuition is that rule variables inferred from different context of the same problem should remain consistent, while the target latent obtained by executing the inferred rule should also be consistent with the latent representation of the ground-truth target. Here, only a subset of the training samples is associated with explicit rule annotations, whereas a much larger set of samples without rule annotations is additionally included. For the latter, the rule induction process is trained solely through the consistency objective. The full-panel teacher may provide corrective signals for both rule induction and target-state inference. The teacher branch is detached from gradient computation only when constructing the OPSD targets. All forward passes associated with the same optimizer step are performed using the same parameter snapshot. Algorithm 1 summarizes one training iteration.

Table 1: Accuracy (%) on RAVEN and I-RAVEN. Each entry reports RAVEN / I-RAVEN.

Models	RAVEN / I-RAVEN							
	Center	L-R	U-D	O-IC	O-IG	2Grid	3Grid	Average
Specialized Methods								
PrAE	14.5/22.6	7.1/21.2	11.1/**26.5**	7.1/16.9	9.5/24.4	13.1/21.4	11.1/18.9	10.5/21.7
LGPP	9.2/20.1	4.7/18.9	5.2/21.2	4.0/13.9	3.1/12.3	8.6/13.7	10.4/13.9	6.5/16.3
ResNet+DRT	14.1/13.2	11.9/13.4	12.8/12.1	13.6/12.1	13.1/13.3	16.8/12.4	16.1/12.8	14.1/12.8
LEN	69.3/15.3	74.5/14.6	74.2/15.5	72.8/12.8	**77.6**/15.7	65.0/15.1	73.5/16.1	72.4/15.0
LVLM-based Methods								
InternVL3-8B	16.0/15.0	6.0/14.0	12.0/13.0	11.0/13.0	17.0/14.0	13.0/11.0	6.0/17.0	11.6/13.9
+Instruct	12.0/16.0	12.0/10.0	11.0/11.0	12.0/11.0	16.0/15.0	10.0/10.0	8.0/18.0	11.6/13.0
GPT-4o	16.0/13.0	13.0/13.0	10.0/6.0	13.0/11.0	10.0/17.0	8.0/13.0	11.0/12.0	11.6/12.1
LLaVA-NeXT	14.0/14.0	13.0/13.0	20.0/17.0	10.0/9.0	10.0/12.0	9.0/15.0	13.0/12.0	12.7/13.1
DeepSeek-VL2	18.0/17.0	22.0/15.0	11.0/11.0	21.0/7.0	20.0/15.0	17.0/11.0	10.0/11.0	17.0/12.4
InternVL3.5-8B	16.0/15.0	6.0/14.0	12.0/13.0	11.0/13.0	17.0/14.0	13.0/11.0	6.0/17.0	11.6/13.9
Qwen3VL-8B	39.4/20.8	26.3/13.4	29.4/16.5	42.5/16.1	40.0/15.1	41.0/15.0	42.3/14.6	37.2/15.9
Qwen3VL-8B+CRD	38.9/41.6	48.1/21.1	50.8/22.8	50.7/29.7	60.3/23.2	**66.2**/25.9	**88.6**/17.5	57.7/28.3
Ours								
LRR	**75.4/53.8**	**83.2/24.2**	**87.3**/26.3	**80.8/34.4**	68.8/**26.0**	63.6/**29.0**	64.8/**25.9**	**74.8/31.3**

4 EXPERIMENTS

Baselines. We compare LRR with representative methods from *Specialized Models* and *LVLM-based Models*. Specialized models include LEN (Zheng et al., 2019), ResNet+DRT (Zhang et al., 2019), LGPP (Shi et al., 2021), and PrAE (Zhang et al., 2021a). For LVLM-based models, we compare against InternVl (Chen et al., 2024c), GPT-4o (Achiam et al., 2023), LLaVA-NeXT (Liu et al., 2024a), DeepSeek-VL2 (Wu et al., 2024), Qwen-VL (Bai et al., 2025a). We also compare against CRD (Shi et al., 2026b), which combines LVLM-based concept extraction with external probabilistic rule modeling. On general visual understanding benchmarks, we additionally compare against general-purpose latent reasoning methods, including CoVT (Qin et al., 2025), LVR (Li et al., 2026), and Monet (Wang et al., 2026). For a fair comparison, we implement LRR and all LVLM-based baselines on the same Qwen3-VL backbone (Bai et al., 2025a).

Benchmarks. We evaluate LRR on three benchmark categories. *In-Distribution Abstract Visual Reasoning Tasks* include RAVEN (Zhang et al., 2019), I-RAVEN (Hu et al., 2021), PGM (Barrett et al., 2018), CVR (Zerroug et al., 2022) and VASR (Bitton et al., 2023). The training splits of these benchmarks are used in LRR training. *Out-of-Distribution Abstract Visual Reasoning Tasks* include MM-IQ (Cai et al., 2025), Do-You-See-Me (Kanade & Ganu, 2026), KiVA (Yiu et al., 2025), MARVEL (Jiang et al., 2024), and VisuLogic (Xu et al., 2026). *Out-of-Distribution General Visual Understanding Tasks* include HRBench4K/8K (Wang et al., 2025b), MMStar (Chen et al., 2024b), MME-RealWorld (Zhang et al., 2025), SpatialEval (Wang et al., 2024), and MMBench (Liu et al., 2024b). More details can be found in Appendix B.

Hyperparameters. We train on $8\times$ A100 GPUs with a batch size of 64. We perform warmup for two epochs and LRSD for one epoch. We use AdamW with learning rates of 1×10^{-5} and 2×10^{-6} for warmup and LRSD. The weight decay is set to 0.01. We use 8 rule latents and 8 execution latents throughout training and inference.

4.1 MAIN RESULTS

In-Distribution Abstract Visual Reasoning. Table 1 shows that LRR achieves the highest average accuracy on RAVEN and I-RAVEN at 74.8% and 31.3%, respectively. On I-RAVEN, it surpasses CRD across all seven configurations and ranks first in six. Table 2 further shows that LRR outperforms the strongest competitor and backbone on each of three additional in-distribution benchmarks.

Table 2: Performance comparison on more abstract visual reasoning benchmarks.

	In Distribution			Out of Distribution				
Models	PGM	VAP	CVR	MM-IQ	VisuLogic	MARVEL	KiVA	Do-You-See-Me
Qwen3-VL-8B	13.9 (0.00)	27.7 (0.00)	31.4 (0.00)	21.8 (0.00)	18.7 (0.00)	30.9 (0.00)	59.9 (0.00)	43.0 (0.00)
+VISER	13.8 (-0.10)	28.8 (+1.10)	32.3 (+0.90)	26.1 (+4.30)	25.4 (+6.70)	31.7 (+0.80)	57.1 (-2.80)	42.7 (-0.30)
+ICoT	13.1 (-0.80)	28.4 (+0.70)	31.5 (+0.10)	24.8 (+3.00)	24.1 (+5.40)	29.1 (-1.80)	60.0 (+0.10)	44.0 (+1.00)
+CoVT	13.7 (-0.20)	28.8 (+1.10)	31.6 (+0.20)	26.0 (+4.20)	24.8 (+6.10)	32.1 (+1.20)	58.0 (-1.90)	45.0 (+2.00)
+LVR	14.5 (+0.60)	26.1 (-1.60)	28.6 (-2.80)	24.8 (+3.00)	24.4 (+5.70)	30.4 (-0.50)	53.6 (-6.30)	43.5 (+0.50)
+Monet	13.8 (-0.10)	26.7 (-1.00)	31.1 (-0.30)	22.8 (+1.00)	15.3 (-3.40)	29.0 (-1.90)	56.1 (-3.80)	42.4 (-0.60)
+CRD	12.5 (-1.40)	27.3 (-0.40)	30.0 (-1.40)	25.9 (+4.10)	23.3 (+4.60)	32.5 (+1.60)	62.6 (+2.70)	44.7 (+1.70)
+LRR	**16.2** (+2.30)	**41.6** (+13.90)	**37.3** (+5.90)	**29.0** (+7.20)	**29.4** (+10.70)	**34.4** (+3.50)	**64.3** (+4.40)	**48.5** (+5.50)

Table 3: Performance comparison on general image understanding tasks.

Models	**Comprehensive**		**RealWorld**			**Spatial**
	MMBench	MMStar	HRBench4K	HRBench8K	MME-RealWorld	SpatialEval
Qwen3-VL-8B	82.96 (0.00)	63.65 (0.00)	74.13 (0.00)	68.50 (0.00)	44.71 (0.00)	64.05 (0.00)
+VISER	84.56 (+1.60)	64.68 (+1.03)	76.75 (+2.62)	70.25 (+1.75)	46.78 (+2.07)	65.74 (+1.69)
+ICoT	83.05 (+0.09)	65.36 (+1.71)	76.63 (+2.50)	70.63 (+2.13)	46.30 (+1.59)	65.48 (+1.43)
+CoVT	84.22 (+1.26)	64.03 (+0.38)	77.00 (+2.87)	70.50 (+2.00)	48.09 (+3.38)	64.61 (+0.56)
+LVR	85.16 (+2.20)	64.89 (+1.24)	77.25 (+3.12)	**72.13** (+3.63)	47.94 (+3.23)	64.97 (+0.92)
+Monet	85.34 (+2.38)	65.28 (+1.63)	77.63 (+3.50)	70.25 (+1.75)	49.04 (+4.33)	65.16 (+1.11)
+LRR	**85.45** (+2.49)	**66.18** (+2.53)	**78.75** (+4.62)	71.88 (+3.38)	**50.47** (+5.76)	**70.80** (+6.75)

These results support the effectiveness of organizing latent reasoning around rule induction and execution across diverse abstract visual reasoning tasks.

Out-of-Distribution Abstract Visual Reasoning. Table 2 shows that LRR achieves the best performance across all five out-of-distribution benchmarks, reaching 29.0% on MM-IQ and 29.4% on VisuLogic. It also outperforms the strongest competitors on MARVEL, KiVA, and Do-You-See-Me. LRR infers relations from each problem's observations and applies them under the query conditions to form a target-state representation. These consistent gains support the transferability of this structured process of latent rule induction and execution beyond the training distribution.

General Visual Understanding. Table 3 shows that LRR improves over Qwen3-VL-8B on all six benchmarks and achieves the best results on five. It surpasses the strongest competing methods by 5.06% on SpatialEval, while ranking second on HRBench8K, only 0.25% below the best result. These gains suggest that latent rule induction and execution can complement visual perception and support reasoning about real-world and spatial relations beyond abstract visual tasks.

4.2 Ablation and Analysis

Ablation of the Main Components and Loss Functions. Table 4 shows that full LRR outperforms standard SFT and all ablated variants across five benchmarks. Removing warmup causes the largest degradation, while omitting the rule prediction loss, target prediction loss, or rule consistency also reduces performance. These results support the complementary roles of initial grounding and subsequent consistency learning in latent rule induction and execution.

Number of Latent Tokens. Table 5 shows that disabling latent tokens $(0 + 0)$ yields performance close to the backbone baseline. All tested nonzero budgets improve performance on most benchmarks, highlighting the benefit of latent rule reasoning. However, these gains diminish when the token budget becomes too large, particularly on RAVEN. Longer latent sequences may introduce redundancy and allow errors to accumulate rather than providing useful additional reasoning steps. We therefore use eight rule tokens and eight target tokens by default as a trade-off between reasoning performance and computational cost.

Table 4: Component Ablation of LRR.

Model	RAVEN	PGM	MM-IQ	KiVA	MME-RW
Baseline					
Qwen3-VL-8B	37.2	13.9	21.8	59.9	44.7
+ Standard SFT	42.5	14.3	23.1	60.4	45.2
Our Method					
LRR	**74.8**	**16.2**	**29.0**	**64.3**	**50.5**
w/o Warmup	38.0	12.4	26.3	50.6	44.9
w/o $\mathcal{J}_{\text{RP}}$	43.7	14.8	26.5	59.7	45.9
w/o $\mathcal{K}_{\text{TP}}$	44.3	15.2	26.6	59.5	47.0
w/o $\mathcal{K}_{\text{RC}}$	72.0	15.0	27.8	63.4	48.3

Table 5: Ablation of Rule and Target Latent.

Rule+Target	RAVEN	PGM	MM-IQ	KiVA	MME-RW
Baseline					
Qwen3-VL-8B	37.2	13.9	21.8	59.9	44.7
LRR					
0+0	37.8	14.1	21.8	60.1	44.9
4+4	**74.8**	14.2	27.9	64.4	49.9
8+8 (default)	**74.8**	16.2	**29.0**	64.3	**50.5**
16+16	74.7	**16.8**	28.0	**65.1**	50.2
32+32	62.4	13.3	26.8	63.8	50.1
64+64	60.4	13.6	27.2	61.5	49.6

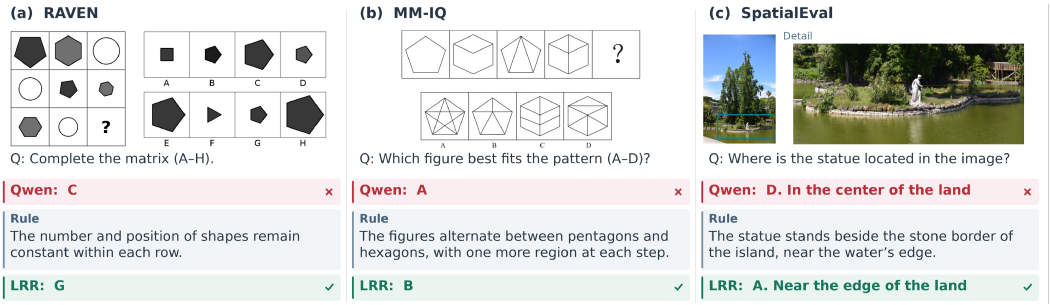

Figure 4: **Qualitative comparison.** Each column shows the visual input, an abbreviated question, the baseline prediction, a rule description, and the LRR prediction.

Comparison across Different Model Scales. Table 6 shows that LRR improves performance across all five benchmarks at both model scales. Notably, the 4B model with LRR surpasses the unmodified 8B backbone on four of the five benchmarks. On RAVEN, LRR yields a gain of 37.6% with the 8B backbone, compared with 5.4% with the 4B backbone.

Table 6: Comparison across model scales.

Model	RAVEN	PGM	MM-IQ	KiVA	MME-RW
Qwen3-VL-8B	37.2	13.9	21.8	59.9	44.7
+LRR	**74.8**	**16.2**	**29.0**	**64.3**	**50.5**
Qwen3-VL-4B	36.7	12.7	20.6	56.2	41.3
+LRR	**42.1**	**15.6**	**26.8**	**58.5**	**46.5**

Case Study. Figure 4 presents qualitative comparisons between LRR and the backbone. On RAVEN and MM-IQ, LRR correctly completes the matrix and continues the sequence where the backbone fails, illustrating the value of inferring relations across observations and applying them to missing elements. On SpatialEval, LRR correctly locates the statue near the land's edge rather than at its center. Together, these cases illustrate how latent rule induction and execution can support both abstract pattern completion and reasoning about relations in natural scenes.

5 Conclusion

We presented Latent Rule Reasoning (LRR), a framework that integrates rule induction and execution within an LVLM's continuous latent space. By connecting inferred relations with their predicted visual consequences, LRR allows rule learning to directly shape the model's internal reasoning process. We further introduced Latent Rule Self-Distillation (LRSD), which uses complete observations during training to guide rule induction under inference conditions, supporting joint learning from examples with and without explicit rule annotations. Extensive experiments demonstrate improved performance in both abstract and general visual reasoning, highlighting the value of latent rule modeling beyond structured abstract tasks. These findings suggest that visual reasoning can benefit not only from recognizing observed content, but also from discovering rules and executing them in latent space to infer unobserved states, thereby enhancing visual reasoning capabilities.

AI USE STATEMENT

In preparing this manuscript, we used AI tools to assist with writing and language polishing, including improving grammar, wording, clarity, and organization. The authors reviewed all AI-assisted text, checking technical descriptions, mathematical notation, and statements about experimental results for accuracy and consistency with the underlying work. We take full responsibility for the final content of this paper, including all text, claims, and artifacts produced with the assistance of AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the LRR architecture and training objectives in Sections 3, with pseudocode provided in Algorithm 1. Section 4 and Appendix B documents the evaluation benchmarks, comparison methods, and training and evaluation settings, including the trainable components, default hyperparameters, and decoding configuration.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Kian Ahrabian, Zhivar Sourati, Kexuan Sun, Jiarui Zhang, Yifan Jiang, Fred Morstatter, and Jay Pujara. The curious case of nonverbal abstract reasoning with multi-modal large language models. *arXiv preprint arXiv:2401.12117*, 2024.

Ahmad Mustafa Anis, Hasnain Ali, and Saquib Sarfraz. On the limitations of vision-language models in understanding image transforms. *arXiv preprint arXiv:2503.09837*, 2025.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In *International conference on machine learning*, pp. 511–520. PMLR, 2018.

Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. *Advances in neural information processing systems*, 29, 2016.

Yonatan Bitton, Ron Yosef, Eliyahu Strugo, Dafna Shahaf, Roy Schwartz, and Gabriel Stanovsky. Vasr: Visual analogies of situation recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 241–249, 2023.

Huanqia Cai, Yijun Yang, and Winston Hu. Mm-iq: Benchmarking human-like abstraction and reasoning in multimodal models. *arXiv preprint arXiv:2502.00698*, 2025.

Patricia A Carpenter, Marcel A Just, and Peter Shell. What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test. *Psychological review*, 97 (3):404, 1990.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024b.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV*, pp. 370–387, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pp. 24185–24198, 2024c.

Hristos S Courellis, Juri Minxha, Araceli R Cardenas, Daniel L Kimmel, Chrystal M Reed, Taufik A Valiante, C Daniel Salzman, Adam N Mamelak, Stefano Fusi, and Ueli Rutishauser. Abstract representations emerge in human hippocampal neurons during inference. *Nature*, 632(8026): 841–849, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Evelina Fedorenko, Steven T Piantadosi, and Edward AF Gibson. Language is primarily a tool for communication rather than thought. *Nature*, 630(8017):575–586, 2024.

Richard Hartley, Andrew Zisserman, et al. *Multiple view geometry in computer vision*, volume 2. Cambridge university press Cambridge, 2003.

Sheng Hu, Yuqing Ma, Xianglong Liu, Yanlu Wei, and Shihao Bai. Stratified rule-aware network for abstract visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1567–1574, 2021.

Yifan Jiang, Jiarui Zhang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski, and Jay Pujara. Marvel: Multidimensional abstraction and reasoning through visual evaluation and learning. *Advances in Neural Information Processing Systems*, 37:46567–46592, 2024.

Aditya Sanjiv Kanade and Tanuja Ganu. Do you see me: A multidimensional benchmark for evaluating visual perception in multimodal llms. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7285–7326, 2026.

Charles Kemp and Joshua B Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692, 2008.

Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Yuxuan Liang, Xu Li, Xiaolei Chen, Haotian Chen, Yi Zheng, Chenghang Lai, Bin Li, and Xiangyang Xue. Global semantic-guided sub-image feature weight allocation in high-resolution large vision-language models, 2025. URL https://arxiv.org/abs/2501.14276.

Yuxuan Liang, Xu Li, Xiaolei Chen, Haotian Chen, Yi Zhen, Zhe Liu, Rui Zhu, Bin Li, and Xiangyang Xue. Pyramid token pruning for high-resolution large vision-language models via region, token, and instruction-guided importance. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2026. doi: 10.1109/TCSVT.2026.3663658.

Daniel R Little, Stephan Lewandowsky, and Thomas L Griffiths. A bayesian model of rule induction in raven's progressive matrices. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34, 2012.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, pp. 34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Yunzhe Liu, Raymond J Dolan, Zeb Kurth-Nelson, and Timothy EJ Behrens. Human replay spontaneously reorganizes experience. *Cell*, 178(3):640–652, 2019.

Andrew Lovett and Kenneth Forbus. Modeling visual problem solving as analogical reasoning. *Psychological review*, 124(1):60, 2017.

Andrew Lovett, Kenneth Forbus, and Jeffrey Usher. A structure-mapping model of raven's progressive matrices. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2010.

Gary F Marcus, Sugumaran Vijayan, Shoba Bandi Rao, and Peter M Vishton. Rule learning by seven-month-old infants. *Science*, 283(5398):77–80, 1999.

Michael McCloskey. Intuitive physics. *Scientific american*, 248(4):122–131, 1983.

Niv Pekar, Yaniv Benny, and Lior Wolf. Generating correct answers for progressive matrices intelligence tests. *arXiv preprint arXiv:2011.00496*, 2020.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*, 2025.

John C Raven and JH Court. *Raven's progressive matrices*. Western Psychological Services Los Angeles, CA, 1938.

Fan Shi, Bin Li, and Xiangyang Xue. Raven's progressive matrices completion with latent gaussian process priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9612–9620, 2021.

Fan Shi, Bin Li, and Xiangyang Xue. Compositional law parsing with latent random functions. In *The Eleventh International Conference on Learning Representations*, 2023.

Fan Shi, Bin Li, and Xiangyang Xue. Towards generative abstract reasoning: Completing raven's progressive matrix via rule abstraction and selection. In *The Twelfth International Conference on Learning Representations*, 2024.

Fan Shi, Haiyang Yu, Bin Li, and Xiangyang Xue. Cola: Chinese character decomposition with compositional latent components. *arXiv preprint arXiv:2506.03798*, 2025.

Fan Shi, Bin Li, and Xiangyang Xue. Abstracting concept-changing rules for solving raven's progressive matrix problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–12, 2026a. doi: 10.1109/TPAMI.2026.3656670.

Fan Shi, Yuxuan Liang, Xiaolei Chen, Haiyang Yu, Xu Li, Yi Zheng, Rui Zhu, Xiangyang Xue, and Bin Li. Decomposition of concept-level rules in visual scenes. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=huEYU44Ax4.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2024.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond images and language. In *CVPR*, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025b.

Yikun Wang, Siyin Wang, Qinyuan Cheng, Zhaoye Fei, Liang Ding, Qipeng Guo, Dacheng Tao, and Xipeng Qiu. Visuothink: Empowering lvlm reasoning with multimodal tree search. *arXiv preprint arXiv:2504.09130*, 2025c.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 25966–26003, 2026.

Weijie Yin, Yongjie Ye, Fangxun Shu, Yue Liao, Zijian Kang, Hongyuan Dong, Haiyang Yu, Dingkang Yang, Jiacong Wang, Han Wang, et al. Sail-vl2 technical report. *arXiv preprint arXiv:2509.14033*, 2025.

Eunice Yiu, Maan Qraitem, Anisa Majhi, Charlie Wong, Yutong Bai, Shiry Ginosar, Alison Gopnik, and Kate Saenko. Kiva: Kid-inspired visual analogies for testing large multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 26900–26919, 2025.

Haiyang Yu, Mengyang Zhao, Jinghui Lu, Ke Niu, Yanjie Wang, Weijie Yin, Weitao Jia, Teng Fu, Yang Liu, Jun Liu, et al. Eve: Towards end-to-end video subtitle extraction with vision-language models. *arXiv preprint arXiv:2503.04058*, 2025.

Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Musslick, and Thomas Serre. A benchmark for compositional visual reasoning. *Advances in neural information processing systems*, 35: 29776–29788, 2022.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5317–5327, 2019.

Chi Zhang, Baoxiong Jia, Song-Chun Zhu, and Yixin Zhu. Abstract spatial-temporal reasoning via probabilistic abduction and execution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9736–9746, 2021a.

Chi Zhang, Sirui Xie, Baoxiong Jia, Ying Nian Wu, Song-Chun Zhu, and Yixin Zhu. Learning algebraic representation for systematic generalization in abstract reasoning. *arXiv preprint arXiv:2111.12990*, 2021b.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. MME-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *The Thirteenth International Conference on Learning Representations*, 2025.

Kecheng Zheng, Zheng-Jun Zha, and Wei Wei. Abstract reasoning with distracting features. *Advances in Neural Information Processing Systems*, 32, 2019.

A Derivations and Proofs

This appendix derives the probabilistic objective and its relationship to the deterministic training procedure. We distinguish exact distributional identities from approximations used by the algorithm. In particular, the forward variational KL uses teacher-distributed prefixes, and evaluating its integrand along a deterministic mean trajectory is an approximation to the corresponding expectation. The analysis below follows the teacher-prefix objectives specified by the training equations.

A.1 Notation and the latent generative model

Fix one training example with visible context $C = (I, T)$, answer A, true target panel U, and, when available, a complete rule annotation R. Write $C^+ = g(C, U)$ for the complete view obtained by restoring U at the query position while retaining the query marker. The construction g supplies privileged visual information during training, which does not insert the answer index or rule text into the student context. In a multiple-choice example, U may be selected from the candidates using the training label A. Let $r = Z^R = (r_1, \dots, r_M)$ and $e = Z^E = (e_1, \dots, e_K)$, where every token lies in $\mathbb{R}^d$, and let $z = (r, e)$. We denote the student means by

$$\mu_m^R(C, r_{<m}) = \mathrm{LVLM}(C, r_{<m}), \quad \mu_k^E(C, r, e_{<k}) = \mathrm{LVLM}(C, r, e_{<k}). \tag{13}$$

Control tokens, segment boundaries, and positional encodings are implicit in this notation. Since the final hidden state and the input embedding have the same dimension d, feeding a hidden state back into the next continuous position is dimensionally well-defined. A control-token boundary uses its ordinary token embedding and advances the position in the usual way. The Gaussian kernels are

$$\begin{aligned} p(r_m \mid C, r_{<m}) &= \mathcal{N}\left(r_m; \mu_m^R(C, r_{<m}), \tau_R^2 I\right), \\ p(e_k \mid C, r, e_{<k}) &= \mathcal{N}\left(e_k; \mu_k^E(C, r, e_{<k}), \tau_E^2 I\right), \\ p(r \mid C) = \prod_{m=1}^{M} p(r_m \mid C, r_{<m}), &\quad p(e \mid C, r) = \prod_{k=1}^{K} p(e_k \mid C, r, e_{<k}). \end{aligned} \tag{14}$$

Throughout the exact derivations, $\tau_R > 0$ and $\tau_E > 0$ are fixed. The answer and rule decoders are normalized distributions.

A.2 Variational Lower Bound

For an annotated example and a fixed $\lambda \geq 0$, define the nonnegative latent factor and its integral by

$$f_\lambda(z) = p(A \mid C, z)\, p(R \mid r)^\lambda\, p(z \mid C), \quad \mathcal{Z}_\lambda(C, A, R) = \int f_\lambda(z)\, \mathrm{d}z. \tag{15}$$

When the rule annotation is absent, the rule factor is omitted and we write $\mathcal{Z}_0(C, A)$. In particular, there is no need to evaluate an undefined rule likelihood and multiply its logarithm by zero. Assume $0 < \mathcal{Z}_\lambda < \infty$. The positive-variance Gaussian variational family has full support on latent space, so the following importance-ratio identity applies:

$$\mathcal{Z}_\lambda = \int q(z \mid C, U) \frac{f_\lambda(z)}{q(z \mid C, U)}\, \mathrm{d}z. \tag{16}$$

Proposition 1 (Weighted ELBO). For any variational distribution,

$$\begin{aligned}\log \mathcal{Z}_\lambda \geq \mathcal{B}_\lambda(q) &= \mathbb{E}_q\left[\log p(A \mid C, z)\right] \\ &\quad + \lambda \mathbb{E}_q\left[\log p(R \mid r)\right] - D_{\mathrm{KL}}\left(q(z \mid C, U) \| p(z \mid C)\right).\end{aligned} \tag{17}$$

The quantity $\mathcal{B}_\lambda$ is the objective denoted $\mathcal{L}_{\mathrm{ELBO}}$ in the main text and is *maximized*.

Proof. Apply Jensen's inequality to the concave logarithm in Equation 16:

$$\begin{aligned}\log \mathcal{Z}_\lambda &= \log \mathbb{E}_q\left[\frac{f_\lambda(z)}{q(z \mid C, U)}\right] \\ &\geq \mathbb{E}_q\left[\log f_\lambda(z) - \log q(z \mid C, U)\right] \\ &= \mathbb{E}_q \log p(A \mid C, z) + \lambda \mathbb{E}_q \log p(R \mid r) - \mathbb{E}_q\left[\log \frac{q(z \mid C, U)}{p(z \mid C)}\right].\end{aligned} \tag{18}$$

The last expectation is the KL divergence in Equation 17. □

Proposition 2 (Exact Variational Gap). Let the normalized tilted latent posterior be

$$\pi_\lambda(z \mid C, A, R) = \frac{f_\lambda(z)}{\mathcal{Z}_\lambda(C, A, R)}. \tag{19}$$

Then

$$\log \mathcal{Z}_\lambda - \mathcal{B}_\lambda = D_{\mathrm{KL}}\left(q(z \mid C, U) \| \pi_\lambda(z \mid C, A, R)\right). \tag{20}$$

Proof. Substituting Equation 19 into the KL gives

$$D_{\mathrm{KL}}(q \| \pi_\lambda) = \mathbb{E}_q[\log q - \log f_\lambda + \log \mathcal{Z}_\lambda] = \log \mathcal{Z}_\lambda - \mathcal{B}_\lambda. \tag{21}$$

Thus equality holds if and only if $q = \pi_\lambda$ almost everywhere. The restriction that q only accesses (C, U) may prevent this equality. In particular, a variational distribution need not use every observed variable, including R, to yield a valid bound. □

Meaning of the rule weight. For $\lambda = 1$, the additional decoder specifies the normalized joint model

$$p(A, R, z \mid C) = p(A \mid C, z) p(R \mid r) p(z \mid C), \tag{22}$$

and $\mathcal{Z}_1(C, A, R) = p(A, R \mid C)$. For $\lambda = 0$, omitting the rule factor gives $\mathcal{Z}_0(C, A) = p(A \mid C)$. For other positive weights, $\mathcal{Z}_\lambda$ is a weighted evidence and is not generally normalized over (A, R). If the sums are finite,

$$\sum_{A,R} \mathcal{Z}_\lambda(C, A, R) = \int p(z \mid C) \sum_R p(R \mid r)^\lambda \, \mathrm{d}z. \tag{23}$$

A separate normalization over (A, R) would change the objective. The variational inequality above bounds the stated weighted evidence without claiming that it is an ordinary likelihood for every λ.

A.3 Exact decomposition into four supervision terms

For readability, abbreviate $q^R = q(r \mid C^+)$, $q^E = q(e \mid U)$, $p^R = p(r \mid C)$, and $p^E = p(e \mid C, r)$.

Proposition 3 (Joint KL Decomposition). The factorized variational family satisfies

$$D_{\mathrm{KL}}(q^R q^E \| p^R p^E) = D_{\mathrm{KL}}(q^R \| p^R) + \mathbb{E}_{q^R}\left[D_{\mathrm{KL}}(q^E \| p^E)\right]. \tag{24}$$

Proof. Expand the log density ratio and integrate each term:

$$\begin{aligned}D_{\mathrm{KL}}(q^R q^E \| p^R p^E) &= \int q^R(r) q^E(e) \left[\log \frac{q^R(r)}{p^R(r)} + \log \frac{q^E(e)}{p^E(e \mid r)}\right] \mathrm{d}e \, \mathrm{d}r \\ &= \int q^R(r) \log \frac{q^R(r)}{p^R(r)} \, \mathrm{d}r \\ &\quad + \int q^R(r) \left[\int q^E(e) \log \frac{q^E(e)}{p^E(e \mid r)} \, \mathrm{d}e\right] \mathrm{d}r.\end{aligned} \tag{25}$$

This is Equation 24. No generative independence of r and e was used. □

Since the rule decoder does not depend on e,

$$\mathbb{E}_{q^R q^E}[\log p(R \mid r)] = \mathbb{E}_{q^R}[\log p(R \mid r)]. \tag{26}$$

Substitution into the ELBO yields

$$\begin{aligned} \mathcal{B}_\lambda &= \mathcal{J}_{\mathrm{AP}} + \lambda \mathcal{J}_{\mathrm{RP}} - \mathcal{K}_{\mathrm{RC}} - \mathcal{K}_{\mathrm{TP}}, \\ \mathcal{J}_{\mathrm{AP}} &= \mathbb{E}_{q^R q^E} \log p_\theta(A \mid C, r, e), \\ \mathcal{J}_{\mathrm{RP}} &= \mathbb{E}_{q^R} \log p_\theta(R \mid r), \\ \mathcal{K}_{\mathrm{RC}} &= D_{\mathrm{KL}}(q^R \| p^R), \\ \mathcal{K}_{\mathrm{TP}} &= \mathbb{E}_{q^R} D_{\mathrm{KL}}(q^E \| p^E). \end{aligned} \tag{27}$$

The loss to minimize is $-\mathcal{B}_\lambda$. Both KL terms are nonnegative. They vanish precisely when $q^R = p^R$ almost everywhere and $q^E = p^E(\cdot \mid r)$ for q^R-almost every r, respectively.

A.4 Autoregressive KL and the common-prefix requirement

Lemma 1 (Autoregressive KL Chain Rule). Let $q(x_{1:n}) = \prod_{j=1}^n q_j(x_j \mid x_{<j})$ and $p(x_{1:n}) = \prod_{j=1}^n p_j(x_j \mid x_{<j})$. Then

$$D_{\mathrm{KL}}(q\|p) = \sum_{j=1}^n \mathbb{E}_{q(x_{<j})} D_{\mathrm{KL}}\left(q_j(\cdot \mid x_{<j}) \| p_j(\cdot \mid x_{<j})\right). \tag{28}$$

Proof. The log ratio factorizes as

$$\log \frac{q(x_{1:n})}{p(x_{1:n})} = \sum_{j=1}^n \log \frac{q_j(x_j \mid x_{<j})}{p_j(x_j \mid x_{<j})}. \tag{29}$$

Take expectation under q. For each summand, integrate out $x_{>j}$ and then condition on $x_{<j}$:

$$\begin{aligned} \mathbb{E}_q \log \frac{q_j(x_j \mid x_{<j})}{p_j(x_j \mid x_{<j})} &= \mathbb{E}_{q(x_{<j})} \mathbb{E}_{q_j(x_j \mid x_{<j})} \log \frac{q_j(x_j \mid x_{<j})}{p_j(x_j \mid x_{<j})} \\ &= \mathbb{E}_{q(x_{<j})} D_{\mathrm{KL}}(q_j(\cdot \mid x_{<j}) \| p_j(\cdot \mid x_{<j})). \end{aligned} \tag{30}$$

Summing proves the result. □

Both conditional distributions in a summand must be evaluated at the *same realized prefix*. For the forward KL, this prefix is distributed according to q, not p. Comparing two means evaluated along separately generated trajectories is not this chain-rule decomposition. Applying the lemma to the rule sequence gives

$$\mathcal{K}_{\mathrm{RC}} = \sum_{m=1}^M \mathbb{E}_{q(r_{<m} \mid C^+)} D_{\mathrm{KL}}\big(q(\cdot \mid C^+, r_{<m}) \| p(\cdot \mid C, r_{<m})\big). \tag{31}$$

For execution, $q(e_k \mid e_{<k}, U) = q(e_k \mid U)$ because the target posterior factorizes across positions. Hence

$$\mathcal{K}_{\mathrm{TP}} = \sum_{k=1}^K \mathbb{E}_{q(r \mid C^+)} \mathbb{E}_{q(e_{<k} \mid U)} D_{\mathrm{KL}}(q(\cdot \mid U) \| p(\cdot \mid C, r, e_{<k})). \tag{32}$$

This expectation includes uncertainty in both the complete rule sequence and the preceding execution tokens.

A.5 Closed-form Solution of Gaussian KL

Lemma 2 (Gaussian KL). For $q = \mathcal{N}(a, \Sigma_q)$ and $p = \mathcal{N}(b, \Sigma_p)$ with positive definite covariances in d dimensions,

$$D_{\mathrm{KL}}(q\|p) = \frac{1}{2}\left[\mathrm{tr}(\Sigma_p^{-1}\Sigma_q) + (a-b)^\top \Sigma_p^{-1}(a-b) - d + \log \frac{\det \Sigma_p}{\det \Sigma_q}\right]. \tag{33}$$

Proof. The log density ratio is

$$\log \frac{q(x)}{p(x)} = \frac{1}{2}\log \frac{\det \Sigma_p}{\det \Sigma_q} + \frac{1}{2}(x-b)^\top \Sigma_p^{-1}(x-b) - \frac{1}{2}(x-a)^\top \Sigma_q^{-1}(x-a). \tag{34}$$

Writing $x - b = (x - a) + (a - b)$ and using $\mathbb{E}_q(x - a) = 0$ gives

$$\begin{aligned} \mathbb{E}_q[(x-b)^\top \Sigma_p^{-1}(x-b)] &= \operatorname{tr}(\Sigma_p^{-1}\Sigma_q) + (a-b)^\top \Sigma_p^{-1}(a-b), \\ \mathbb{E}_q[(x-a)^\top \Sigma_q^{-1}(x-a)] &= d. \end{aligned} \tag{35}$$

Taking expectations proves Equation 33. □

For equal isotropic covariances, all covariance terms cancel exactly:

$$D_{\mathrm{KL}}\big(\mathcal{N}(a, \tau^2 I) \| \mathcal{N}(b, \tau^2 I)\big) = \frac{\|a-b\|_2^2}{2\tau^2}. \tag{36}$$

There is no remaining additive constant in this KL. With unequal isotropic variances σ_q^2 and σ_p^2, the expression instead is

$$\frac{\|a-b\|_2^2}{2\sigma_p^2} + \frac{d}{2}\left(\frac{\sigma_q^2}{\sigma_p^2} - 1 + \log \frac{\sigma_p^2}{\sigma_q^2}\right). \tag{37}$$

Thus equal fixed variances justify the particular pure squared-error reduction used here.

Combining Lemmas 1 and 2 yields the exact expressions

$$\begin{aligned} \mathcal{K}_{\mathrm{RC}} &= \frac{1}{2\tau_R^2} \sum_{m=1}^{M} \mathbb{E}_{q(r_{<m}|C^+)} \left\| \mu_m^R(C, r_{<m}) - \nu_m^R(C^+, r_{<m}) \right\|_2^2, \\ \mathcal{K}_{\mathrm{TP}} &= \frac{1}{2\tau_E^2} \sum_{k=1}^{K} \mathbb{E}_{q(r|C^+)q(e_{<k}|U)} \left\| \mu_k^E(C, r, e_{<k}) - h_k \right\|_2^2. \end{aligned} \tag{38}$$

The integration over the *current* token has been performed analytically. The expectations over its conditioning prefixes remain. In particular, replacing a random prefix by its mean is not an additional consequence of the Gaussian KL identity.

A.6 Answer and rule likelihoods as token cross-entropies

Let $A = (a_1, \ldots, a_{L_A})$ be the answer-token sequence, including the termination token when modeled. By autoregressive factorization,

$$\begin{aligned} p(A \mid C, r, e) &= \prod_{l=1}^{L_A} p(a_l \mid a_{<l}, C, r, e), \\ \ell_A(r, e) &= -\log p(A \mid C, r, e) = \sum_{l=1}^{L_A} \mathrm{CE}\left(p(\cdot \mid a_{<l}, C, r, e), a_l\right). \end{aligned} \tag{39}$$

Here $\mathrm{CE}(\pi, a) = -\log \pi(a)$ for a one-hot target. Therefore $-\mathcal{J}_{\mathrm{AP}} = \mathbb{E}_{q^R q^E} \ell_A$. A candidate classifier replaces the token sum by $-\log p(A = j \mid C, r, e)$ for the correct candidate j. These are alternative answer parameterizations, and a textual-answer token must be supervised by a_l, not by the entire answer sequence at every position. Let $S(R) = (s_1, \ldots, s_{L_R})$ be a fixed, complete serialization of the annotated rule set. Then

$$\begin{aligned} p(R \mid r) = p(S(R) \mid r) &= \prod_{l=1}^{L_R} p(s_l \mid s_{<l}, r), \\ \ell_R(r) = -\log p(R \mid r) &= \sum_{l=1}^{L_R} \mathrm{CE}(p(\cdot \mid s_{<l}, r), s_l), \\ \mathcal{J}_{\mathrm{RP}} &= -\mathbb{E}_{q^R} \ell_R(r). \end{aligned} \tag{40}$$

Teacher forcing supplies ground-truth prefixes of the output text. The only example-specific conditioning representation is r. This likelihood covers all J rules when the serialization retains every attribute-relation binding. The number of latent positions M need not equal the number of rules J. A canonical serialization makes the target unambiguous; invariance to arbitrary rule permutations would require a separate construction and does not follow from this token likelihood.

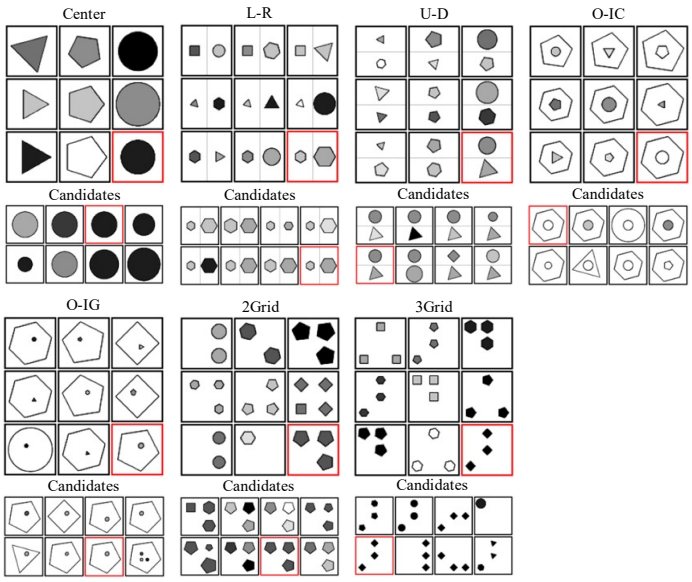

Figure 5: **Visualization of the RAVEN dataset.** Examples illustrate the seven configurations in RAVEN/I-RAVEN, with red borders marking the correct answers that satisfy the rules established by the context panels.

B Detailed Experimental Settings

Appendix B.1 introduces the benchmarks used in our evaluation, and Appendix B.2 describes the compared methods.

B.1 Benchmarks

We evaluate LRR on three groups of benchmarks. *In-Distribution Abstract Visual Reasoning* includes RAVEN (Zhang et al., 2019), I-RAVEN (Hu et al., 2021), PGM (Barrett et al., 2018), and VASR (Bitton et al., 2023), covering matrix completion and visual analogy. *Out-of-Distribution Abstract Visual Reasoning Tasks* include MM-IQ (Cai et al., 2025), Do-You-See-Me (Kanade & Ganu, 2026), KiVA (Yiu et al., 2025), MARVEL (Jiang et al., 2024), and VisuLogic (Xu et al., 2026). This group covers diverse reasoning formats together with perceptual and spatial tasks that probe the visual capabilities underlying abstract reasoning. *Out-of-Distribution General Visual Understanding Tasks* include HRBench (Wang et al., 2025b), MMStar (Chen et al., 2024b), MME-RealWorld (Zhang et al., 2025), SpatialEval (Wang et al., 2024), and MMBench (Liu et al., 2024b). These benchmarks assess fine-grained perception, spatial understanding, and reasoning over real-world visual content. Together, the three groups examine LRR's performance on in-distribution rule-based tasks and its transfer to broader visual reasoning and understanding.

B.1.1 In-Distribution Abstract Visual Reasoning

RAVEN (Zhang et al., 2019). RAVEN contains 70,000 Raven's Progressive Matrices problems across seven visual configurations, including centered objects, grids, spatially separated components, and nested layouts. Each problem presents an incomplete 3×3 matrix and eight answer candidates. The panels follow rules involving attributes such as number, position, type, size, and color, with four rule types: Constant, Progression, Arithmetic, and Distribute Three. An attributed stochastic image grammar generates the visual configurations and their hierarchical structural annotations. Solving a problem requires identifying attribute-level relations across the context panels and selecting the candidate that completes the matrix consistently with those relations.

I-RAVEN (Hu et al., 2021). I-RAVEN, or Impartial-RAVEN, revises RAVEN's answer generation to address an identified shortcut in the original candidate sets. It preserves the context matrices and

underlying rules while reconstructing the answer candidates using an Attribute Bisection Tree. This procedure balances the attribute variations among candidates, counteracting the tendency of the correct answer to be identifiable from attribute frequencies in the answer set alone. The benchmark therefore places greater emphasis on recovering relations from the context panels rather than exploiting this answer-only shortcut, while retaining RAVEN's structured matrix-completion setting.

PGM (Barrett et al., 2018). Procedurally Generated Matrices (PGM) evaluates abstract reasoning with incomplete 3×3 matrices and eight candidate answers. Each problem is generated from relation–object–attribute triples, combining objects such as shapes and lines with attributes such as number, position, size, type, and color. The relation vocabulary includes progression, XOR, OR, AND, and consistent union. These structured generators allow relevant relations and distracting visual variations to be controlled separately. The benchmark provides a neutral regime and additional regimes involving interpolation, extrapolation, and held-out combinations, enabling analysis of how models generalize across attribute values and relational structures.

VASR (Bitton et al., 2023). Visual Analogies of Situation Recognition (VASR) extends visual analogy to real-world scenes. Given a reference image pair and a third image, the task is to select the fourth image that preserves the analogous relationship. The dataset is constructed using situation recognition annotations and includes a human-validated collection of 3,820 analogies. Unlike benchmarks restricted to geometric transformations, VASR requires understanding the situations depicted in natural images and matching their relational structure. Carefully selected distractors make it necessary to distinguish a valid analogy from candidates that are visually or semantically related but do not complete the required relationship.

B.1.2 Out-of-Distribution Abstract Visual Reasoning

MM-IQ (Cai et al., 2025). MM-IQ evaluates multimodal abstraction and reasoning through 2,710 curated questions spanning eight reasoning paradigms: logical operation, mathematics, two-dimensional geometry, three-dimensional geometry, visual instruction, temporal movement, spatial relationship, and concrete object. Its tasks include identifying geometric relationships, interpreting changes in visual patterns, and reasoning about spatial configurations. The benchmark emphasizes reasoning from visual evidence rather than dependence on specialized domain knowledge. Its varied problem formats and reasoning requirements provide a broader assessment than matrix completion alone, testing whether a model can identify and apply the relevant relationships across diverse visual problems.

Do-You-See-Me (Kanade & Ganu, 2026). Do-You-See-Me evaluates visual perception through 1,758 images and 2,612 questions. Its seven subtasks cover shape discrimination, joint shape–color discrimination, letter disambiguation, form constancy, spatial relations, figure–ground separation, and visual closure. Representative tasks include counting specified shapes, recognizing patterns after geometric transformations, locating objects, and matching incomplete shapes to their complete forms. The benchmark uses programmatically generated visual stimuli with controlled difficulty. Rather than exclusively testing rule induction, it probes whether models correctly recover the visual details and spatial structure on which subsequent reasoning depends.

KiVA (Yiu et al., 2025). KiVA evaluates visual analogy using transformations of familiar objects across five domains: color, size, number, rotation, and reflection. Models must infer a transformation from a reference example and apply it to a different object. The base KiVA dataset contains 1,400 object transformations; the separate KiVA-adults extension contains 2,900 and introduces more demanding generalization. Its evaluation distinguishes identifying the changing visual domain, specifying the transformation, and selecting the appropriate transformed target. This separation allows the benchmark to probe both the recognition of a visual rule and its application beyond the particular object used to demonstrate it.

MARVEL (Jiang et al., 2024). MARVEL contains 770 abstract visual reasoning puzzles covering six patterns: temporal movement, spatial relationship, quantities, mathematical operations, two-dimensional geometry, and three-dimensional geometry. The puzzles use both geometric and abstract shapes across five configurations: sequence, two-row, matrix, group, and reassembling. Beyond the final reasoning question, its hierarchical evaluation includes coarse-grained perception questions about the puzzle layout and fine-grained questions about relevant visual details. This design distinguishes recognition of the input structure from inference over the underlying pattern,

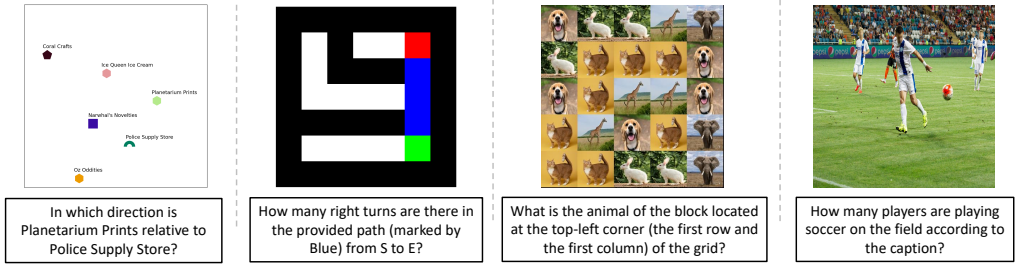

Figure 6: **Visualization of the SpatialEval benchmark.** SpatialEval comprehensively assesses spatial intelligence in LLMs and VLMs across four dimensions: spatial relationships, positional understanding, object counting, and navigation.

allowing analysis of whether an incorrect answer originates from a perceptual error or a failure to reason over correctly perceived content.

VisuLogic (Xu et al., 2026). VisuLogic contains 1,000 human-verified multiple-choice problems organized into six categories: Quantitative, Spatial, Positional, Attribute, Stylistic, and Other. The tasks include reasoning about changes in graphical element counts, folding and unfolding three-dimensional structures, translation and rotation, symmetry, and shape overlays or differences. Its problems are selected to emphasize visual information that may be lost in coarse textual descriptions. Solving them requires preserving fine-grained graphical details while inferring the transformations or constraints that connect the presented figures, providing an assessment of vision-centric reasoning beyond semantic image recognition.

B.1.3 General Visual Understanding

HRBench (Wang et al., 2025b). HRBench evaluates high-resolution image understanding through HRBench4K and HRBench8K. It contains two task groups: Fine-grained Single-instance Perception and Fine-grained Cross-instance Perception. The former assesses details of individual instances, including attributes and textual content, while the latter requires combining information across instances, such as determining relative spatial positions or interpreting maps and charts. The benchmark includes 200 image queries, with 4K images obtained by cropping the corresponding 8K images around relevant objects. It tests whether models can locate small but informative visual regions and use their details within the broader image context.

MMStar (Chen et al., 2024b). MMStar contains 1,500 human-selected questions covering six core capabilities and 18 detailed evaluation axes. The core capabilities are coarse perception, fine-grained perception, instance reasoning, logical reasoning, science and technology, and mathematics. Its construction emphasizes visual dependence, with filtering and manual review intended to reduce questions that can be answered through text alone and to limit potential data leakage. The benchmark combines perceptual recognition with reasoning over image-grounded evidence, providing a compact assessment of whether a model effectively uses visual information across both general and knowledge-related tasks.

MME-RealWorld (Zhang et al., 2025). MME-RealWorld evaluates perception and reasoning in high-resolution, application-oriented visual scenarios. The benchmark comprises 13,366 images and 29,429 manually annotated question–answer pairs across 43 subtasks. Its five domains are OCR in the Wild, Diagram and Table, Remote Sensing, Autonomous Driving, and Monitoring. Questions require capabilities such as reading small text, counting densely distributed objects, identifying spatial relationships, and interpreting structured visual information. The combination of detailed imagery and varied real-world tasks makes it an assessment of whether models can ground their answers in local evidence while understanding the surrounding scene.

SpatialEval (Wang et al., 2024). SpatialEval studies spatial reasoning through four tasks: Spatial-Map, Maze-Nav, Spatial-Grid, and Spatial-Real. These tasks assess spatial relationships, navigation, position understanding, and object counting using both synthetic visual environments and real im-

ages. Representative questions ask about relative directions, turns along a marked route, objects at specified grid locations, and relations between objects in natural scenes. The original benchmark provides text-only, vision-only, and combined vision–text formats to analyze the role of input modality. It therefore distinguishes spatial reasoning from the ability to extract the relevant spatial information from visual observations.

MMBench (Liu et al., 2024b). MMBench is a general multimodal evaluation benchmark with multiple-choice questions and parallel English and Chinese versions. Its hierarchical taxonomy spans perception and reasoning, with six intermediate capability groups and 20 fine-grained abilities. These include coarse perception, single-instance and cross-instance fine-grained perception, attribute reasoning, logical reasoning, and relation reasoning. The original evaluation framework also introduces CircularEval, which rotates candidate order to assess the consistency of answer selection. Its diverse task coverage provides a broad assessment of visual understanding beyond abstract patterns, including recognition of image content and reasoning about attributes and relationships.

B.2 Other Methods

We summarize the compared methods in three groups: *Probabilistic Rule Modeling*, *Discriminative Relational Reasoning*, and *LVLM-Assisted Rule Modeling*. The formulations below describe their original computational procedures and learning objectives under a common notation. For matrix reasoning, $\mathcal{C} = \{I_i\}_{i\in\mathcal{O}}$ denotes the observed panels, $\mathcal{A} = \{A_k\}_{k=1}^{K_A}$ denotes the answer candidates, and y denotes the correct candidate index. The missing panel is denoted by I_*, and z_i denotes a latent visual representation. Additional symbols are defined locally for each method.

B.2.1 Probabilistic Rule Modeling

PrAE (Zhang et al., 2021a) combines neural perception with probabilistic abduction and execution. Object predictions are marginalized into panel-level attribute distributions P_i^a. For attribute a, the probability of a rule r^a is obtained by summing over compatible context assignments:

$$p(r^a \mid \mathcal{C}) \propto \sum_{S^a \in \mathcal{V}(r^a)} \prod_{i\in\mathcal{O}} P_i^a(S_i^a), \tag{41}$$

where $\mathcal{V}(r^a)$ contains assignments satisfying the rule's logical constraints. Execution propagates context distributions under the selected rule to predict the target distribution $\widehat{P}_*^a$. Rules are sampled during training and selected by maximum posterior probability during testing. Candidates are compared using summed Jensen–Shannon divergences:

$$d_k = \sum_a D_{\mathrm{JS}}\left(\widehat{P}_*^a \| P_{A_k}^a\right), \qquad p(k \mid \mathcal{C}, \mathcal{A}, r) = \frac{\exp(-d_k)}{\sum_j \exp(-d_j)}. \tag{42}$$

The minimum-divergence candidate is selected. With r collecting the attribute-specific rules, the learning objective includes answer supervision and an auxiliary rule-prediction loss:

$$\mathcal{L}_{\mathrm{PrAE}} = \mathbb{E}_{r\sim p(r|\mathcal{C})}\left[-\log p(y \mid \mathcal{C}, \mathcal{A}, r)\right] - \sum_a \lambda_a \log p(r_{\mathrm{gt}}^a \mid \mathcal{C}). \tag{43}$$

Here, r_{gt}^a is the annotated rule for attribute a. REINFORCE handles discrete rule selection, and curriculum learning stabilizes perception and reasoning.

LGPP (Shi et al., 2021) models matrix completion through latent Gaussian process priors. An encoder infers panel representations and latent row and column coordinates. Each latent dimension follows a separate rule function:

$$z_{ij}^d = f^d(\lambda_i^{\mathrm{row}}, \lambda_j^{\mathrm{col}}), \qquad f^d \sim \mathcal{GP}(0, \kappa_\theta^d). \tag{44}$$

Conditioned on the inferred coordinates and context values, the missing representation follows the noise-free GP predictive distribution:

$$\begin{aligned} p(z_*^d \mid z_{\mathcal{O}}^d, \lambda) &= \mathcal{N}(\mu_*^d, (\sigma_*^d)^2), \\ \mu_*^d &= k_{*\mathcal{O}}^d (K_{\mathcal{O}\mathcal{O}}^d)^{-1} z_{\mathcal{O}}^d, \\ (\sigma_*^d)^2 &= k_{**}^d - k_{*\mathcal{O}}^d (K_{\mathcal{O}\mathcal{O}}^d)^{-1} k_{\mathcal{O}*}^d. \end{aligned} \tag{45}$$

The covariance terms are evaluated using the learned kernels, and a decoder maps z_* to the missing image. Training uses a weighted variational objective. Its marginal-matching approximation is

$$\begin{aligned}\widetilde{\mathcal{J}}_{\mathrm{LGPP}} = \mathbb{E}_q &\left[\sum_{i\in\mathcal{O}\cup\{*\}} \log p_\psi(I_i \mid z_i)\right] \\ &- \beta D_{\mathrm{KL}}(q_\phi(\lambda \mid \mathcal{C}) \| p(\lambda)) \\ &- \gamma \mathbb{E}_{q_\phi(\lambda|\mathcal{C})}\left[\sum_d \sum_{i\in\mathcal{O}} D_{\mathrm{KL}}\big(q_\phi(z_i^d \mid I_i) \| p_\theta(z_i^d \mid \lambda)\big)\right].\end{aligned} \tag{46}$$

Here, q is the joint variational distribution inferred from the context. Marginal matching replaces the intractable joint latent KL term with tractable panel-wise divergences.

CLAP-NP (Shi et al., 2023) represents concept-specific laws using Neural Processes. Context images are encoded into concept representations z_i^c. A separate function parser aggregates the context pairs (u_i, z_i^c), where u_i is the conditioning input, to infer a Gaussian function variable g^c. Target concepts and images are generated through

$$\begin{aligned} g^c &\sim p_\theta(g^c \mid z^c_{\mathcal{O}}, u_{\mathcal{O}}), \\ z_t^c &\sim \mathcal{N}\big(T_p^c(g^c, u_t), \sigma_z^2 \mathbf{I}\big), \\ I_t &\sim p_\psi(I_t \mid [z_t^1; \ldots; z_t^D]). \end{aligned} \tag{47}$$

The regularized variational objective is

$$\mathcal{J}_{\mathrm{CLAP}} = \mathcal{L}_{\mathrm{rec}} - \beta_t \mathcal{K}_{\mathrm{target}} - \beta_f \mathcal{K}_{\mathrm{function}} - \beta_{\mathrm{TC}} \mathcal{R}_{\mathrm{TC}}. \tag{48}$$

Here, $\mathcal{L}_{\mathrm{rec}}$ is the target-image reconstruction log-likelihood. The target KL term matches encoded target concepts to their predicted distributions. The function KL term matches the posterior inferred from context and target observations to the context-conditioned prior. Sharing the context encoder makes the context KL term vanish. Concept decomposition is additionally encouraged through total correlation:

$$\mathcal{R}_{\mathrm{TC}} = D_{\mathrm{KL}}\left(\bar{q}(z) \,\middle\|\, \prod_c \bar{q}(z^c)\right), \tag{49}$$

where $\bar{q}$ denotes the aggregated posterior over dataset images.

B.2.2 Discriminative Relational Reasoning

LEN (Zheng et al., 2019) uses a Logic Embedding Network to evaluate relations among panel triples. Each candidate completes the matrix, and a shared visual encoder produces embeddings e_i. The resulting triples are divided into row/column triples $\mathcal{T}_k^{\parallel}$ and the remaining triples $\mathcal{T}_k^{\mathrm{other}}$. For a 3×3 matrix, these groups contain six and 78 triples, respectively. Two relation networks aggregate their information together with a global context embedding u:

$$s_k = h_\psi\left(\sum_{(i,j,\ell)\in\mathcal{T}_k^{\parallel}} g_{\phi_1}(e_i, e_j, e_\ell, u) + \sum_{(i,j,\ell)\in\mathcal{T}_k^{\mathrm{other}}} g_{\phi_2}(e_i, e_j, e_\ell, u)\right). \tag{50}$$

The basic network is trained with answer classification:

$$\mathcal{L}_{\mathrm{LEN}} = -\log \frac{\exp(s_y)}{\sum_{k=1}^{K_A} \exp(s_k)}, \qquad \hat{y} = \arg\max_k s_k. \tag{51}$$

The original work separately evaluates a reinforcement-learning teacher that controls training-data proportions and an auxiliary type-prediction loss; these are extensions rather than defining components of the basic LEN architecture.

ResNet+DRT (Zhang et al., 2019) augments a ResNet-18 visual backbone with a Dynamic Residual Tree. A serialized structural description S determines the computation tree. Given backbone features

h, each node combines its incoming features with its label embedding e_v:

$$a_v = \begin{cases} h, & v \text{ is a leaf}, \\ \sum_{u \in \mathrm{ch}(v)} h_u, & \text{otherwise}, \end{cases} \qquad h_v = \mathrm{ReLU}\big(f_{\ell(v)}([a_v; e_v])\big), \tag{52}$$

where $\mathrm{ch}(v)$ denotes the children of v, and transformations are shared among nodes at the same tree level $\ell(v)$. The root output forms a residual update before answer classification:

$$\widetilde{h} = h + h_{\mathrm{root}}, \qquad s = \mathrm{MLP}(\widetilde{h}), \qquad \mathcal{L}_{\mathrm{DRT}} = -\log \mathrm{softmax}(s)_y. \tag{53}$$

The original experiments retrieve structural annotations to assemble the trees. Thus, the tree specifies a structure-conditioned neural computation, rather than an inferred symbolic rule program.

SRAN (Hu et al., 2021) learns rule embeddings at cell-wise, individual-wise, and ecological levels, corresponding to individual panels, single rows, and pairs of rows. Gated embedding fusion progressively combines these representations. Averaging outputs after exchanging the two input rows provides row-order invariance. Let H denote the resulting SRAN encoder, M_1, M_2 the complete context rows, and $M_3^{(k)}$ the third row completed by candidate A_k. The reference and candidate rule embeddings are

$$g = H(M_1, M_2), \qquad r_k = \frac{1}{2}\left[H(M_1, M_3^{(k)}) + H(M_2, M_3^{(k)})\right]. \tag{54}$$

Column-wise embeddings are constructed analogously and concatenated with the row-wise embeddings before scoring. Using the resulting embeddings, SRAN applies inner-product similarity and the $(K_A + 1)$-tuplet loss:

$$s_k = g^\top r_k, \qquad \mathcal{L}_{\mathrm{SRAN}} = \log\left(1 + \sum_{k \neq y} \exp(s_k - s_y)\right). \tag{55}$$

Inference selects $\hat{y} = \arg\max_k s_k$, favoring the candidate whose induced rule embedding best matches the reference.

B.2.3 LVLM-Assisted Rule Modeling

CRD (Shi et al., 2026b) uses a pretrained LVLM to propose visual concepts and estimate their values over image patches. For concept c, the patch-value vector v^c is modeled by a zero-mean Gaussian process over patch positions p, with deep-kernel covariance K_ϕ. Its rule-consistency score and kernel-learning objective are

$$\ell_c = \log \mathcal{N}(v^c; 0, K_\phi), \qquad \mathcal{L}_{\mathrm{CRF}} = -\mathbb{E}_{(p, v^c)}\left[\log p_\phi(v^c \mid p)\right]. \tag{56}$$

The kernel parameters are learned while the LVLM remains frozen. A visual concept set G of fixed cardinality K is then sampled according to

$$p_K(G \mid \ell) \propto \exp\left(\sum_{c \in G} \ell_c\right), \qquad |G| = K. \tag{57}$$

The sampler uniformly selects $i \in G$ and uses the LVLM to propose a replacement $j \notin G$. For $G' = G \setminus \{i\} \cup \{j\}$, the Metropolis–Hastings acceptance probability is

$$\alpha(G, G') = \min\left\{1, \exp(\ell_j - \ell_i)\frac{q(i \mid j, G')}{q(j \mid i, G)}\right\}, \tag{58}$$

where q is the LVLM-based proposal distribution. Accepted proposals update the concept set; rejected proposals retain it. For matrix reasoning, each candidate fills the missing panel, and the completed matrix with the highest CRF score determines the answer. The LVLM supplies concept proposals, while the external GP supplies rule-based scoring.

C Limitations and Future Work

Although LRR achieves strong performance across diverse visual reasoning tasks, its fixed latent token budget may not fully accommodate differences in task complexity. Images governed by more complex or interacting rules may benefit from additional latent tokens, whereas simpler tasks may require fewer. A promising direction is therefore to allocate latent tokens adaptively according to the reasoning demands of each input. LRR naturally lends itself to this extension, as the budgets for rule induction and execution could be adjusted separately while preserving their distinct roles, potentially achieving a better balance between reasoning capacity and computational cost.